Iniziato	domenica, 11 settembre 2022, 16:34
Stato	Completato
Terminato	domenica, 11 settembre 2022, 16:34
Tempo impiegato	4 secondi
Valutazione	**0,00** su un massimo di 13,00 (**0**%)

Domanda 1

Risposta non data

Punteggio max.: 4,00

Un sociologo desidera indagare i fattori che influiscono sul benessere fisico e mentale. in un campione di 13 persone tra i 35 e i 50 anni misura un indice che misura se la persona mostra segni di malessere (Y) e un indice che indica il tempo dedicato al lavoro (Z). Il ricercatore desidera indagare qual è l'effetto di Z su Y e usa un modello lineare semplice, verificando che sia significativo rispetto al modello nullo.

```
summary(df)
```

```
       z               y        
 Min.   :0.039   Min.   :57.4  
 1st Qu.:2.184   1st Qu.:62.1  
 Median :3.710   Median :64.6  
 Mean   :4.377   Mean   :63.8  
 3rd Qu.:6.755   3rd Qu.:66.3  
 Max.   :7.684   Max.   :67.8  
```

```
fit <- lm(y~z,data=df)
summary(fit)
```

```

Call:
lm(formula = y ~ z, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-4.912 -2.414  0.685  1.884  4.362 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   61.297      1.632   37.56  5.8e-13 ***
z              0.567      0.324    1.75     0.11    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.9 on 11 degrees of freedom
Multiple R-squared:  0.217, Adjusted R-squared:  0.146 
F-statistic: 3.05 on 1 and 11 DF,  p-value: 0.109
```

Il parametro di interesse è il coefficiente angolare (β_1) che descrive l'effetto di Z su Y. In prima istanza il ricercatore desidera calcolare un intervallo di confidenza per il coefficiente angolare (β_1) e fare un test statistico che testi il seguente sistema di ipotesi:

$$H_0 : \beta_1 = 0.92 \quad VS \quad H_1 : \beta_1 \neq 0.92$$

Infine il ricercatore calcola degli intervalli di confidenza per il valore atteso del livello di malessere per i valori del predittore z, $z_0 = 4$ e $z^* = 9$.

```
# CI per z_0
predict(fit, newdata = data.frame(z = 4), level = 1-alpha, interval = "confidence")
```

```
    fit   lwr   upr
1 63.56 61.67 65.46
```

```
# CI per z^*
predict(fit, newdata = data.frame(z = 9), level = 1-alpha, interval = "confidence")
```

```
  fit lwr upr
1  NA  NA  NA
```

Si indichi il valore del limite superiore dell'intervallo di confidenze per il coefficiente angolare (β_1). Si indichi poi il valore della statistica test per il sistema di ipotesi di interesse. Si indichi se il modello stimato risulta significativo rispetto al modello nullo. Infine si indichi quale degli intervalli forniti è l'intervallo di confidenza trovato dalla funzione `predict` per $z^* = 9$

Il livello di significatività usato per test e intervalli di confidenze è: $\alpha = 0.04$.

a. Si indichi il limite superiore dell'intervallo di confidenza per il coefficiente angolare β_1.

✖

b. Si indichi il valore della statistica test per il test su β_1 = 0.92.

✖

c. Il modello è significativo rispetto al modello nullo.

Il modello non è significativo rispetto al modello nullo.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: Il modello non è significativo rispetto al modello nullo.

d. CI(z^*): (62.55, 64.57).

CI(z^*): (62.75, 64.37).

CI(z^*): (62.43, 70.36).

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: CI(z^*): (62.43, 70.36).

L'intervallo di confidenza per il valore del coefficiente angolare β_1 si trova usando la seguente formula:

$$\hat{\beta} \pm t_{\alpha/2,n-2} * sd(\hat{\beta})$$

Il valori $\hat{\beta}$ e $sd(\hat{\beta})$ sono disponibili nell'output della funzione summary: $\hat{\beta}$ = 0.5666 e $sd(\hat{\beta})$ = 0.3245. Infine $t_{\alpha/2,n-2}$ = `qt(0.98, 11)` = 2.328:

$$0.5666 + 0.3245 * 2.328 = 1.322$$

La statistica test per il sistema di ipotesi indicato è:

$$tstat = \frac{\hat{\beta} - \beta_1^0}{sd(\hat{\beta})} = \frac{0.5666 - 0.92}{0.3245} = -1.0891$$

Per verificare se il modello `fit` è significativo, si deve controllare il valore del pvalue associato alla statistica F nel summary: si dice che un modello è significativo rispetto al modello nullo se il p-value è minore di 0.04.

Il valore z^* è al di fuori del range dei valori osservati per il predittore Z: l'intervallo di confidenza stimato per questo valore del predittore sarà molto ampio.

a. 1.32
b. -1.09
c. Falso / Vero
d. Falso / Falso / Vero

Domanda 2

Risposta non data

Punteggio max.: 6,00

L'ufficio commerciale di una catena di supermercati desidera indagare la relazione tra il numero di dipendenti e il fatturato nei diversi punti vendita della catena. Vengono quindi raccolte informazioni per l'anno passato sul fatturato medio mensile in centinaia di migliaia di euro (indicato con la variabile F), il numero di dipendenti (indicato con la variabile D), e la posizione del punto vendita (P). Quest'ultima variabile è categoriale è può prendere tre valori: `altro`, `centro_citta` e `periferia`.

Il primo modello stimato usa entrambe le variabili che vengono inserite in maniera additiva:

```
fitLoc <- lm(f~d+p,data=df)
summary(fitLoc)$coefficient
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      99.45     1.8120   54.88 1.01e-68
d                 0.43     0.0243   17.66 2.40e-30
pcentro_citta    -5.67     1.6112   -3.52 6.91e-04
pperiferia       14.78     1.7306    8.54 4.19e-13
```

```
confint(fitLoc, parm = "d", level = 0.96)
```

```
    2 % 98 %
d 0.379 0.48
```

```
qt(p = c(0.98,0.96), df = fitLoc$df.residual)
```

```
[1] 2.09 1.77
```

L'analista costruisce un test per verificare l'ipotesi che il coefficiente angolare β_1 nel modello, che indica l'effetto della variabile d su f sia uguale a 0.5, cioè per verificare il seguente test di ipotesi

$$H_0 : \beta_1 = 0.5 \quad VS \quad H_1 : \beta_1 \neq 0.5$$

Deriva poi una stima puntuale e intervallare per il fatturato di un determinato punto vendita (Punto A) con le seguenti caratteristiche:

```
puntoA
```

```
   d           p
1 21 centro_citta
```

```
coef(fitLoc)
```

```
  (Intercept)             d pcentro_citta    pperiferia
        99.45          0.43         -5.67         14.78
```

```
predict(fitLoc, newdata = puntoA, se.fit = TRUE)
```

```
$fit
[1] NA

$se.fit
[1] 1.3

$df
[1] 86

$residual.scale
[1] 6.3
```

Infine, tramite un test anova, verifica se dai dati vi sia un'indicazione che l'effetto del numero dei dipendenti sul fatturato vari per i punti vendita in diverse posizioni:

```
fitLocInt <- lm(f~d*p,data=df)
anova(fitLoc, fitLocInt)
```

```
 Analysis of Variance Table

 Model 1: f ~ d + p
 Model 2: f ~ d * p
   Res.Df  RSS Df Sum of Sq    F Pr(>F)
 1     86 3408
 2     84 3064  2       344 4.72  0.011 *
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si indichi la statistica test per l'ipotesi $H_0 : \beta_1 = 0.5$ e si indichi se l'ipotesi può essere rifiutata. Si indichi in quale tipo di punto vendita si registra un fatturato più alto a parità di numero di dipendenti. Si indichino poi la stima puntuale e il valore del limite superiore dell'intervallo di confidenza per il fatturato del punto di vendita A. Infine si indichi se vi è un'indicazione che l'effetto del numero dei dipendenti sul fatturato vari per i punti vendita in diverse posizioni.

Per tutti i calcoli si usi un valore di significatività $\alpha = 0.04$.

a. La statistica test per l'ipotesi $H_0 : \beta_1 = 0.5$:.

✖

b. ○ L'ipotesi nulla non può essere rifiutata.

○ Non è possibile stabilire se l'ipotesi può essere rifiutata.

○ L'ipotesi nulla può essere rifiutata.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: L'ipotesi nulla può essere rifiutata.

c. ○ A parità di dipendenti il fatturato più alto in media è nei negozi in centro città.

○ A parità di dipendenti il fatturato più alto in media è nei negozi in periferia.

○ A parità di dipendenti il fatturato più alto in media è nei negozi nella categoria altro.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: A parità di dipendenti il fatturato più alto in media è nei negozi in periferia.

d. Il valore stimato per il punto vendita A:.

✖

e. Il limite superiore per l'intervallo di confidenza per il punto A:.

✖

f. ○ Non vi è un'indicazione che l'effetto del numero dei dipendenti sul fatturato sia lo stesso per i punti vendita in diverse posizioni.

○ Vi è un'indicazione che l'effetto del numero dei dipendenti sul fatturato sia lo stesso per i punti vendita in diverse posizioni.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: Non vi è un'indicazione che l'effetto del numero dei dipendenti sul fatturato sia lo stesso per i punti vendita in diverse posizioni.

La statistica test si deriva con la formula:

$$\frac{\hat{\beta} - 0.5}{sd(\hat{\beta})} = \frac{0.43 - 0.5}{0.024} = -2.888$$

Si può poi rifiutare H_0 se il valore assoluto della statistica test appena calcolata è maggiore del percentile della distribuzione T: 2.085. In alternativa si può anche guardare se il valore 0.5 è all'interno dell'intervallo di confidenza: se questo avviene non si può rifiutare H_0.

Il modello stimato è:

$$\text{Per negozi in categoria altro: } f = 99.448 + 0.43 * d$$

$$\text{Per negozi in centro: } f = (99.448 + (-5.673)) + 0.43 * d$$

$$\text{Per negozi in periferia: } f = (99.448 + (14.785)) + 0.43 * d$$

La stima puntuale per il punto A è quindi:

$$f = (99.448 + (-5.673) + (0)) + 0.43 * 21 = 102.799$$

Il limite superiore dell'intervallo di confidenza si può derivare con:

$$102.799 + 2.085 * 1.303$$

Dato che il p-value del test ANOVAè minore di 0.04, non vi è un'indicazione che l'effetto del numero dei dipendenti sul fatturato sia lo stesso per i punti vendita in diverse posizioni. Il test infatti verifica se tutti i coefficienti che descrivono una diversa dipendenza da d per ogni tipo di punto venidta siano pari a zero.

a. -2.89
b. Falso / Falso / Vero
c. Falso / Vero / Falso
d. 102.80
e. 105.52
f. Vero / Falso

Domanda 3

Risposta non data

Punteggio max.: 3,00

Un sociologo desidera indagare i fattori che influiscono sul benessere fisico e mentale. In un campione di 55 persone tra i 35 e i 50 anni misura un indice che misura se la persona mostra segni di malessere (Y), un indice che indica il tempo dedicato al lavoro (Z) e due indici legati al tempo dedicato agli spostamenti casa-lavoro (X1 e X2). In prima istanza il ricercatore desidera indagare qual è l'effetto di Z su Y e se questo è significativo.

```
fitZ <- lm(y~z)
summary(fitZ)
```

```
Call:
lm(formula = y ~ z)

Residuals:
   Min     1Q Median     3Q    Max
-13.12  -5.41  -1.31   7.08  13.93

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  123.204      2.103    58.6   <2e-16 ***
z              0.589      0.422     1.4       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.61 on 53 degrees of freedom
Multiple R-squared:  0.0355,    Adjusted R-squared:  0.0173
F-statistic: 1.95 on 1 and 53 DF,  p-value: 0.168
```

Il ricercatore stima poi un modello in cui anche la variabile X2 è inserita nel modello e desidera capire se aggiungere questa variabile aumenta in maniera significativa la bontà di adattamento del modello:

```
Analysis of Variance Table

Model 1: y ~ z
Model 2: y ~ z + x2
  Res.Df  RSS Df Sum of Sq   F Pr(>F)
1     53 3070
2     52  243  1      2827 606 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Infine il ricercatore vuole valutare quale modello ha un AIC più basso tra un modello in cui la variabile Z e X1 vengono incluse come predittori o un modello in cui la variabili Z e X2 vengono incluse come predittori.

```
fit1 <- lm(y~z+x1)
```

```
logLik(fit1)
```

```
'log Lik.' -188.58 (df=4)
```

```
logLik(fit2)
```

```
'log Lik.' -118.86 (df=4)
```

Nell'analisi viene usato un livello di significatività di $\alpha = 0.05$.

Si indichi il valore del p-value mancante nel summary. Si indichi poi se il la bontà di adattamento del modello `fit2` è significativamente migliore della bontà di adattamento del modello `fitZ`. Si indichi infine quale tra i modelli `fit1` e `fit2` ha un AIC maggiore.

a. Si indichi il p-value mancante nel `summary`.

✖

b. ○ La bontà di adattamento del modello fit2 è significativamente migliore di quella del modello fitZ.

○ La bontà di adattamento del modello fit2 non è significativamente migliore di quella del modello fitZ.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: La bontà di adattamento del modello fit2 è significativamente migliore di quella del modello fitZ.

c. ○ AIC(fit2) < AIC(fit1).

○ AIC(fit2) > AIC(fit1).

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: AIC(fit2) < AIC(fit1).

Il pvalue mancante si calcola a partire dalla statistica test mostrata nel summary: `2*pt(abs(tstat),df=n-1,lower.tail=FALSE) =pt(1.39625,df=54,lower.tail=FALSE)=0.16846`

Dato che il p-value nella tabella prodotta da `anova` è minore di 0.05 si può affermare che la bontà di adattamento del modello `fit2` è migliore di quella del modello `fitZ`.

Il due modelli `fit1` e `fit2` hanno lo stesso numero di parametri: il modello che ha verosimiglianza minore avrà anche un valore di AIC maggiore, dato che

$$AIC(M(p)) = -2 * logLik(M) + 2 * p.$$

a. 0.17
b. Vero / Falso
c. Vero / Falso

◄ Vecchi appelli - esercizi sul modello di regressione semplice

Vai a...

Vecchi appelli - esercizi sui GLM ►